

# NLP在美团点评的应用

美团点评-智能技术中心-自然语言处理

李彪

2017.06.17



## Adversarial Generation of Natural Language

Yoav Goldberg



Yann LeCun



# 目录

01 NLP在美团点评的应用场景

02 NLP在美团点评的应用技术

03 NLP在美团点评的应用案例

04 总结

# 01 NLP在美团点评的应用场景

文本：Query

NLP: Query分析  
如分词，成分分析，意图识别，改写...

文本：商家标题详情

NLP: 商家详情分析  
如品类，地址等识别





# 01 NLP在美团点评的应用场景

NLP: 评价Tag抽取

文本: 用户评论数据

NLP: 评论分类

如: 垃圾, 黄反, 优质等识别

NLP: 情感分析

如: 正向, 负向, 中性



# 01 NLP在美团点评的应用场景

NLP: 头条的分类

文本: 头条新闻

NLP:标题生成、内容理解



# 01 NLP在美团点评的应用场景



文本： 客服对话内容

```
Thread 0 Crashed:
ImageIO      _CGImageReadSessionFinalize(void const*) + 44
CoreFoundation  _CFRelease + 216
ImageIO      IIOImagePlus::~IIOImagePlus() + 168
ImageIO      IIOImagePlus::~IIOImagePlus() + 20
ImageIO      _CGImagePlusFinalize(void const*) + 192
CoreFoundation  _CFRelease + 216
CoreFoundation  _CFBasicHashDrain + 288
CoreFoundation  _CFRelease + 216
ImageIO      IIOReadPlugin::ReleaseInfo(void*) + 32
ImageIO      IIOImageProviderInfo::~IIOImageProviderInfo() + 156
ImageIO      IIOImageProviderInfo::~IIOImageProviderInfo() + 20
ImageIO      IIOImageProviderInfo::ReleaseInfo(void*) + 164
CoreGraphics    image_provider_finalize + 148
CoreFoundation  _CFRelease + 216
CoreGraphics    image_finalize + 88
CoreFoundation  _CFRelease + 216
```

.....

文本： Crash日志

.....

## 02 NLP在美团点评的应用技术

NLP研究什么问题？

“自然语言处理可以定义为研究在人与人交际中以及在人与机器交际中的语言问题的一门学科”-马纳瑞斯



## 02 NLP在美团点评的应用技术

### NLP研究什么问题？

词法短语：分词，词性标注，命名实体识别，组块分析，Term权重，Term紧密度

句法语义：语言模型，依存句法分析，词义消歧，语义角色标注，深层语义分析

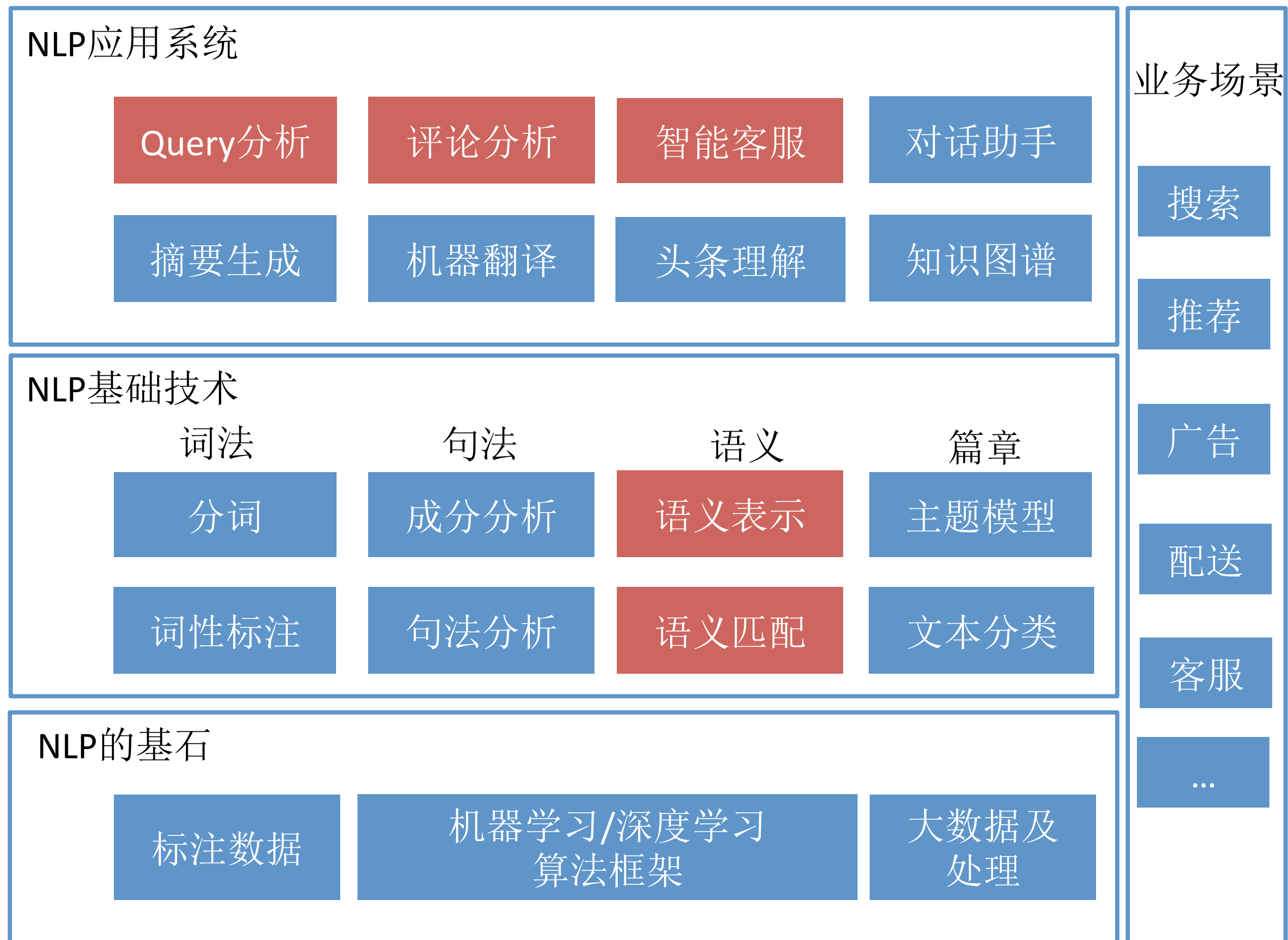
篇章理解：文本分类、聚类，文章摘要，文本生成，篇章关系识别，篇章衔接关系，指代消歧，语义表示，语义匹配，主题模型，情感分析，舆情监控

系统应用：信息抽取，知识图谱（表示，建图，补全，推理等）  
信息检索（索引，召回，排序等），Query分析，自动问答，  
智能对话，阅读理解，机器翻译，语音识别，合成，  
OCR, 图像文字生成

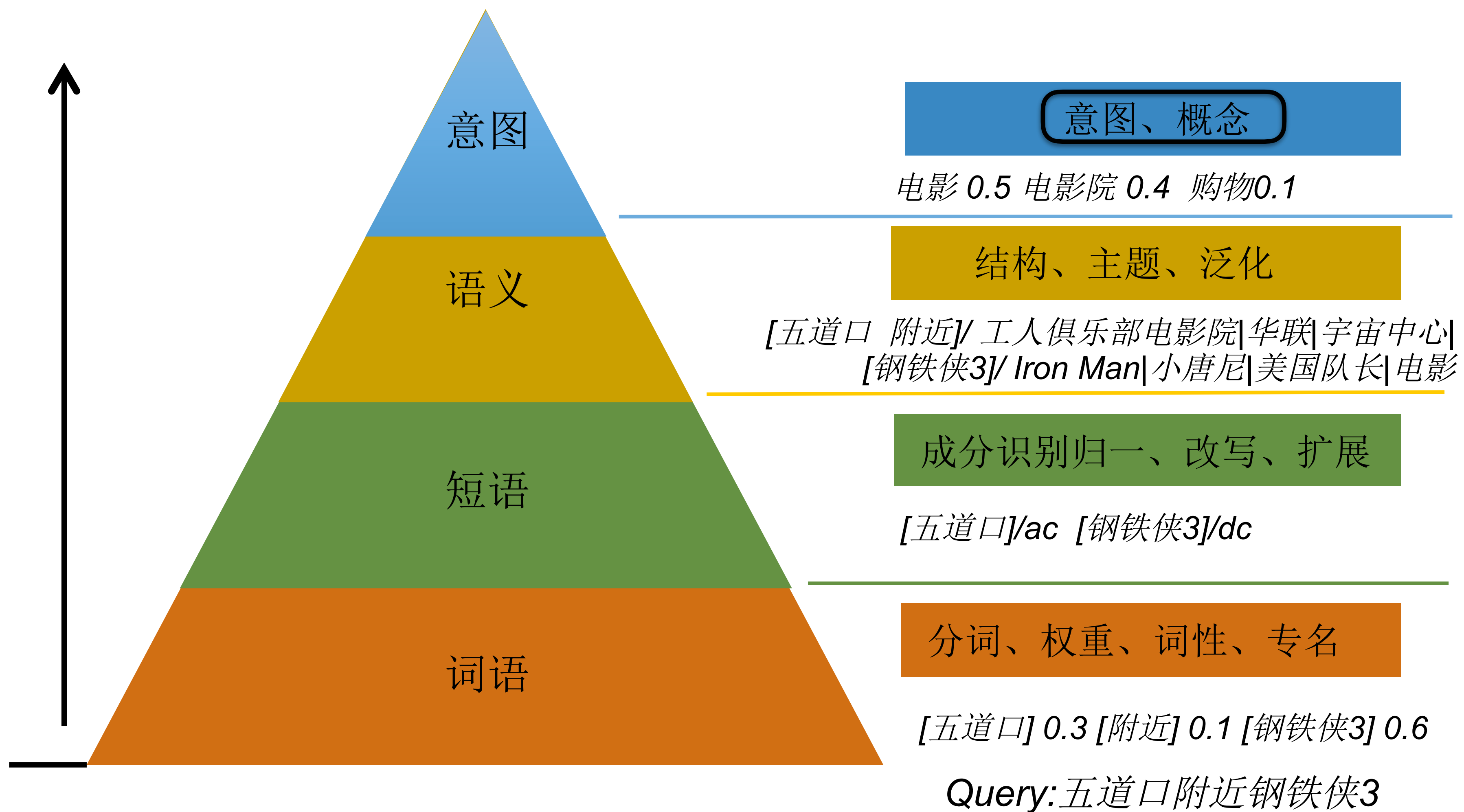
.....

## 02 NLP在美团点评的应用技术

### NLP@美团点评



# 案例1-Query意图识别

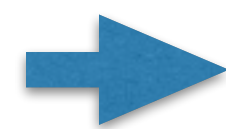


# 案例1-Query意图识别

## 场景

- Q:动物园门票 ➡ 意图:旅游
- Q:全聚德望京 ➡ 意图:美食

Query





# 案例1-Query意图识别

## 思路

问题转换：Query意图识别->短文本分类问题->机器学习多分类

处理流程：



# 案例1-Query意图识别

## 思路

### 1 训练样本获取

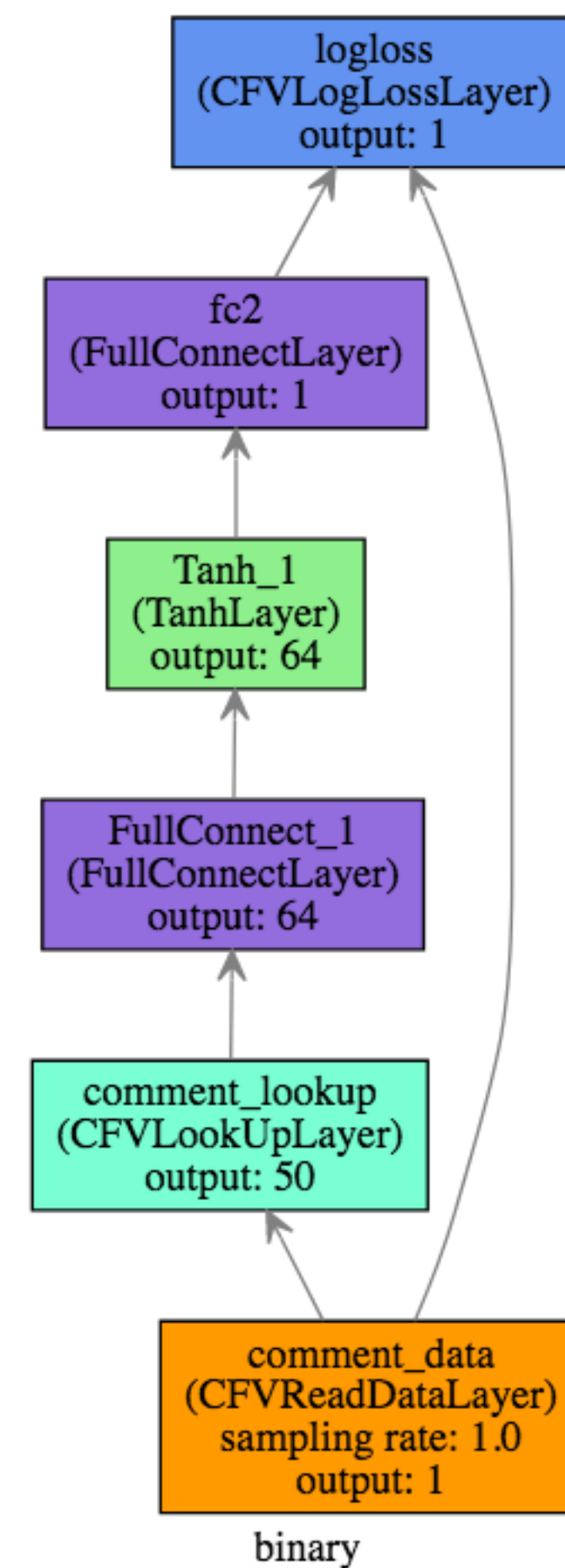
- 自动标注：统计值，主动学习，Bootstrapping,
- 人工标注

### 2 特征设计特点

- Query语义表示
- Query的信息扩展
- 用户地域：eg 黄鹤楼

### 3 模型尝试

- Xgboost
- DNN on Ginger



DNN on Ginger 文本分类模型

# 案例1-Query意图识别

## 问题分析

- 分类目标
- 品类体系
  - 分发 召回 排序
- 现状分析
- 词典策略分发
  - 分发P 95+ R 50+/-
- 数据支持
- 分发词典
- 解决方案
- ML

## 数据准备

- 数据标注
- 正例\_SVM  
行为信息+分发词典
  - 正例\_DNN  
用户行为
  - 采样方式
    - 下采样

## 特征工程

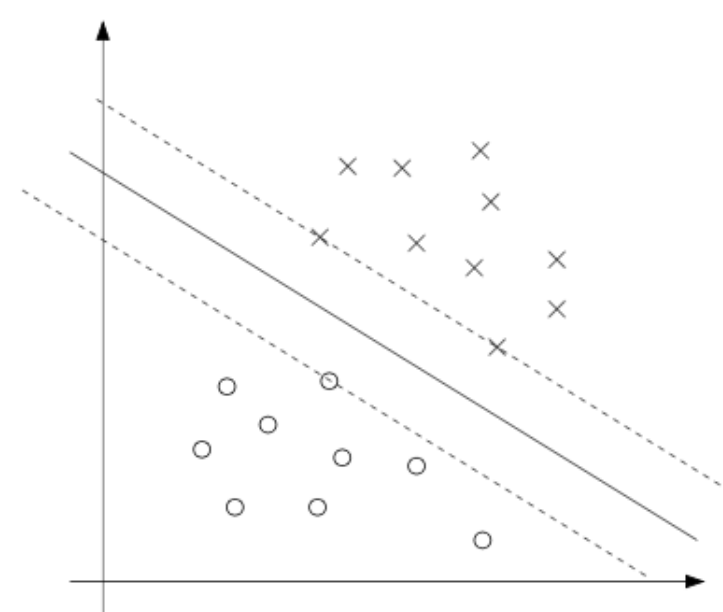
- 特征选择\_SVM
- 词法+句法
  - 任务特征
    - 垂类资源匹配度
    - 加权点击概率
    - 垂类关键词
- 特征选择\_DNN
- 词向量
  - 城市

## 模型训练

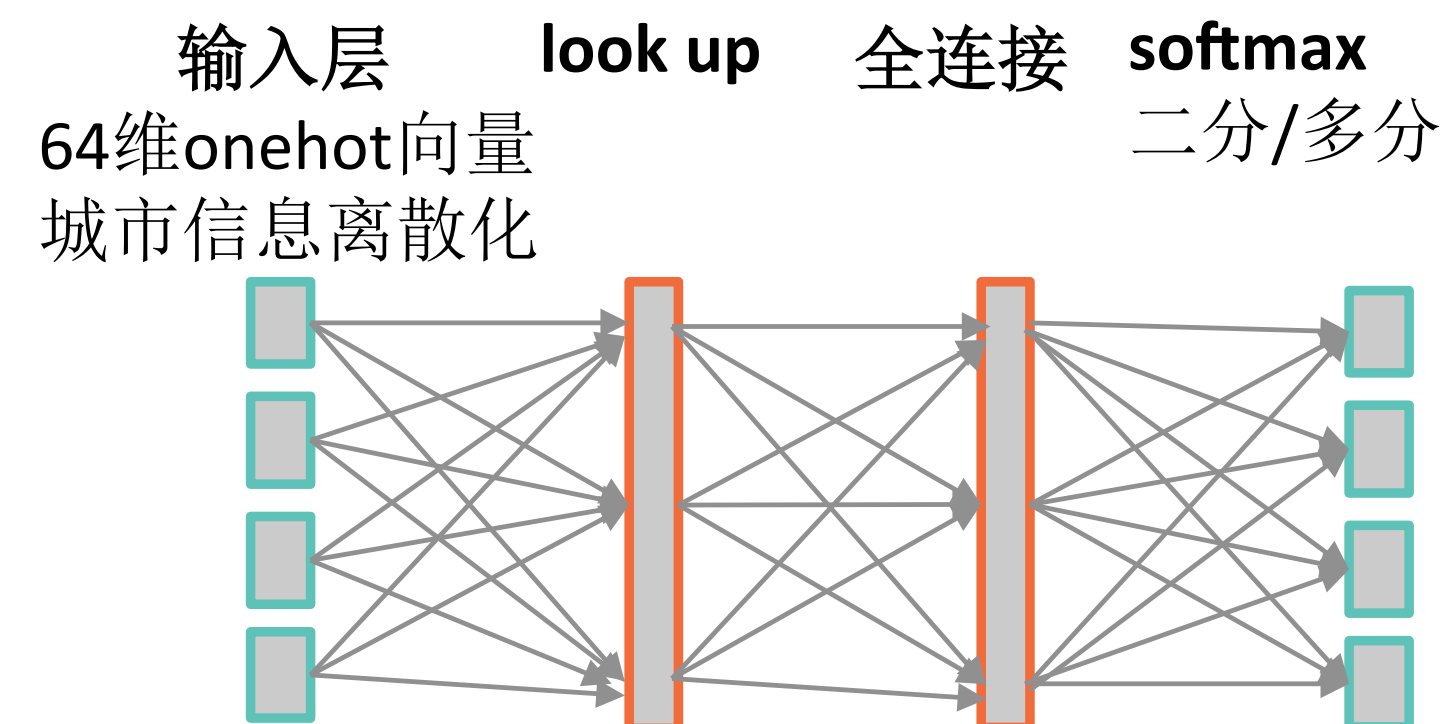
- 模型选择
- SVM xgboost
  - DNN

## 效果评估

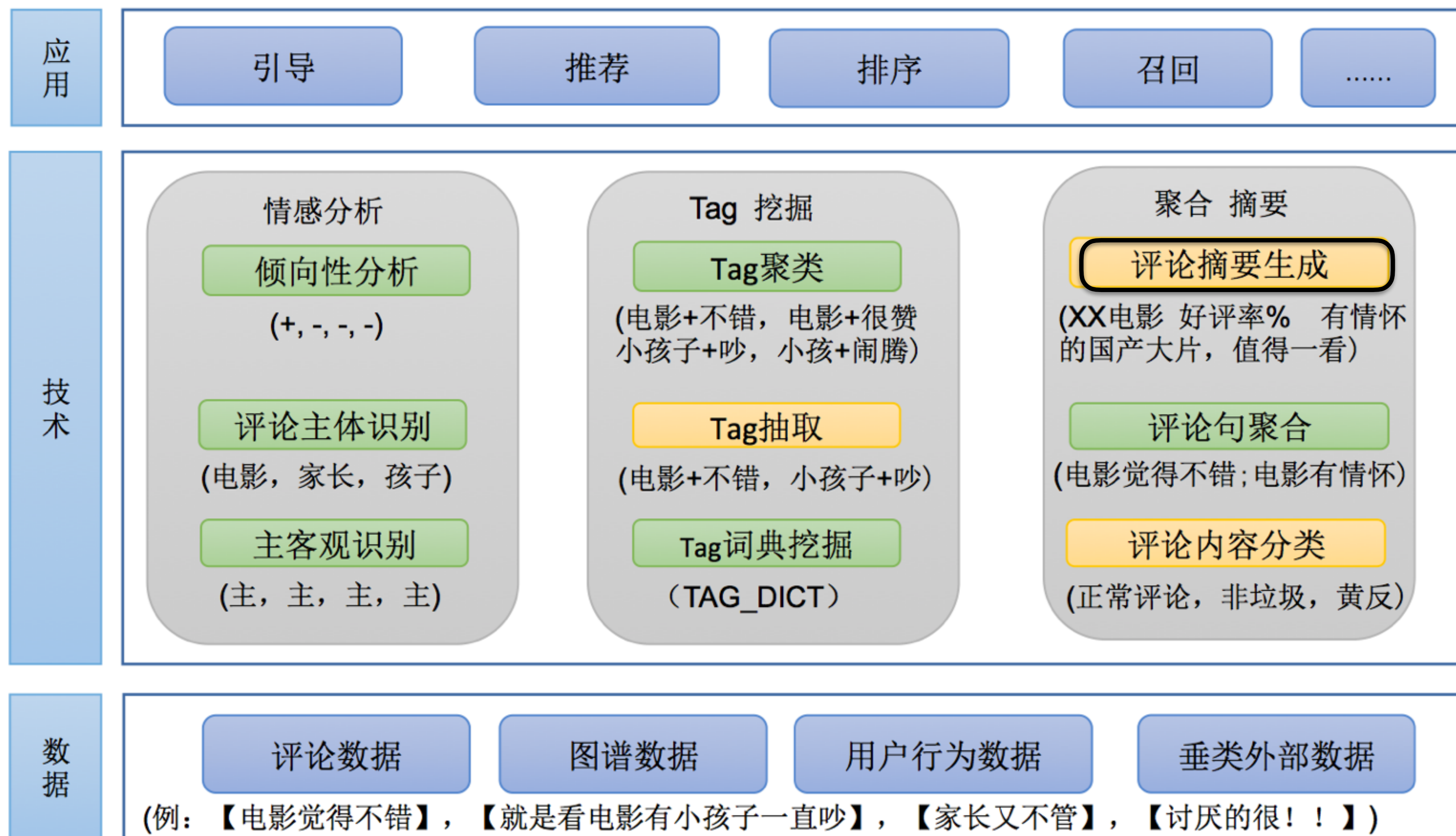
- 线下评估
- F\_SVM 旅游 85 酒店95
  - F\_DNN +2
- 线上评估(规则+模型)
- 旅游场景 相对提升5.05%
  - 酒店场景 相对提升5.19%



词典依赖  
规则依赖  
特征稀疏



## 案例2 商家评价摘要



线上直接应用

间接应用



# 案例2 商家评价摘要

场景

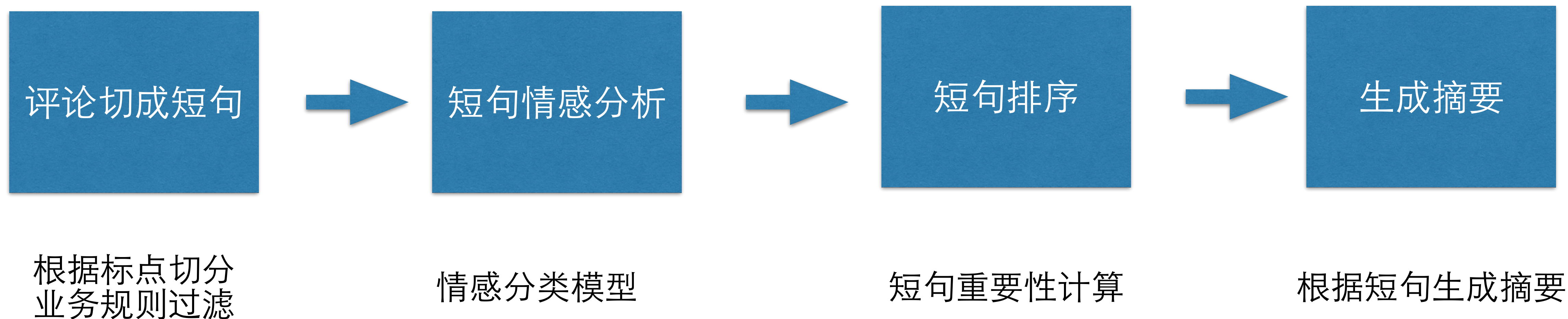
来源：评论，商家详情，菜品描述等  
难点：摘要生成问题



掌柜说：海鲜划算，觉得味道不错，店里服务好

## 案例2 商家评价摘要

思路



已在推荐猜你喜欢，和广告掌柜说上线

# 案例3 语义匹配模型

## 场景

对象：

Query-POI

Query-Query

POI-POI

业务：

1 搜索，广告等

召回，排序系统

2 POI去重，

Query推荐，

Query扩展等





## 案例3 语义匹配模型

### 经典方法

#### 向量空间

- 1970左右
- TF-IDF, BM25...
- 评价：优- 简单实用；缺- 同义词，多义词，维度大

#### 矩阵分解

- 1990左右
- LSA..
- 评价： 优-解决维度大，同义词问题， 缺-物理含义弱

#### 概率图模型

- 2000年左右
- pLSA (99'), LDA(03'),
- 评价：数学完备，物理含义明确

#### 深度学习

- 2013-
- Word2vec(13'), DSSM(13'), ClickNet..
- 评价：语义，语法



## 案例3 语义匹配模型

Ours

特点:

表示层:

语义特征表示

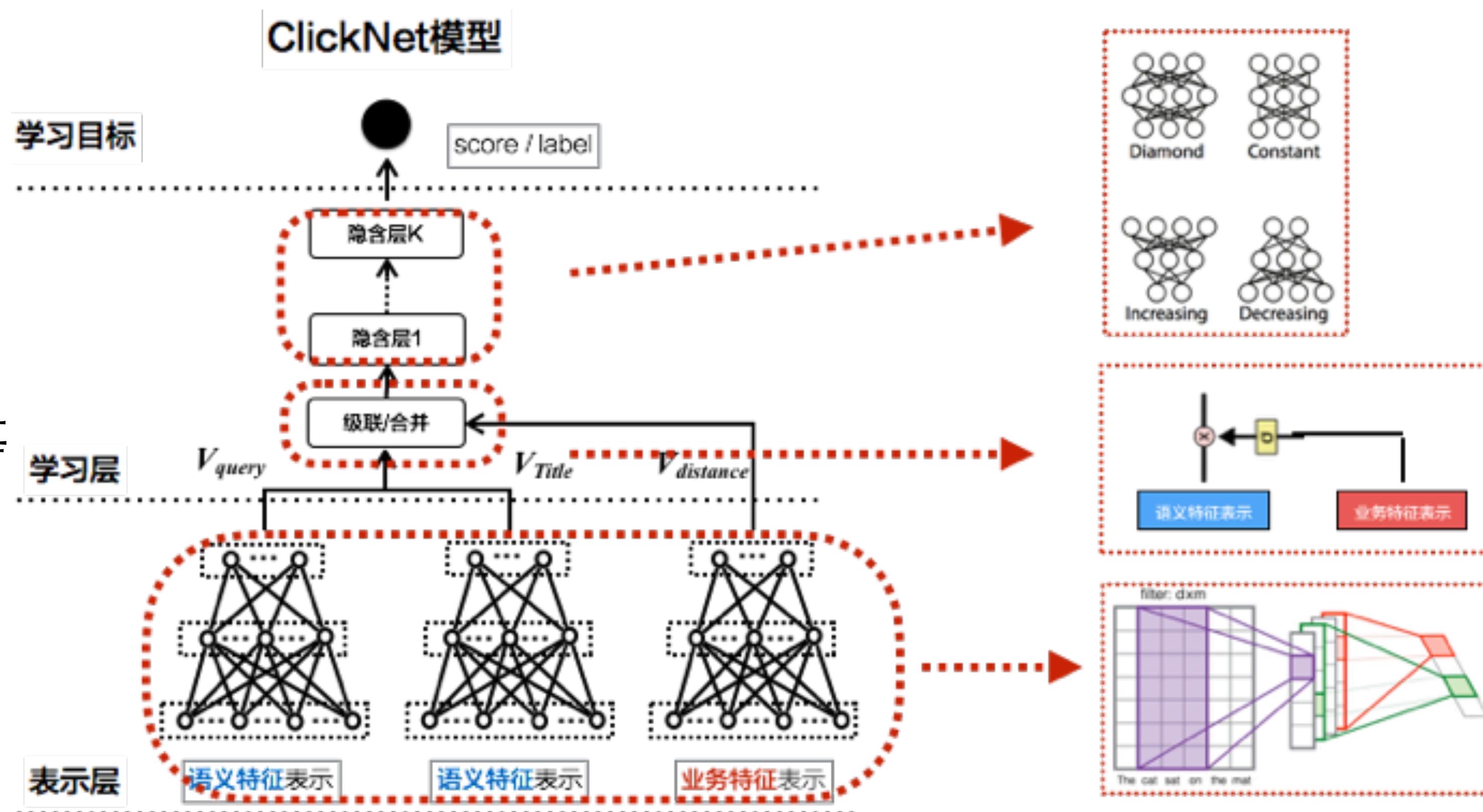
比如句子 BOW, CNN, RNN

业务特征表示

比如poi-query距离、用户评价等

学习层:

信息融合，多层非线性变化



ClickNet框架

## 案例3 语义匹配模型

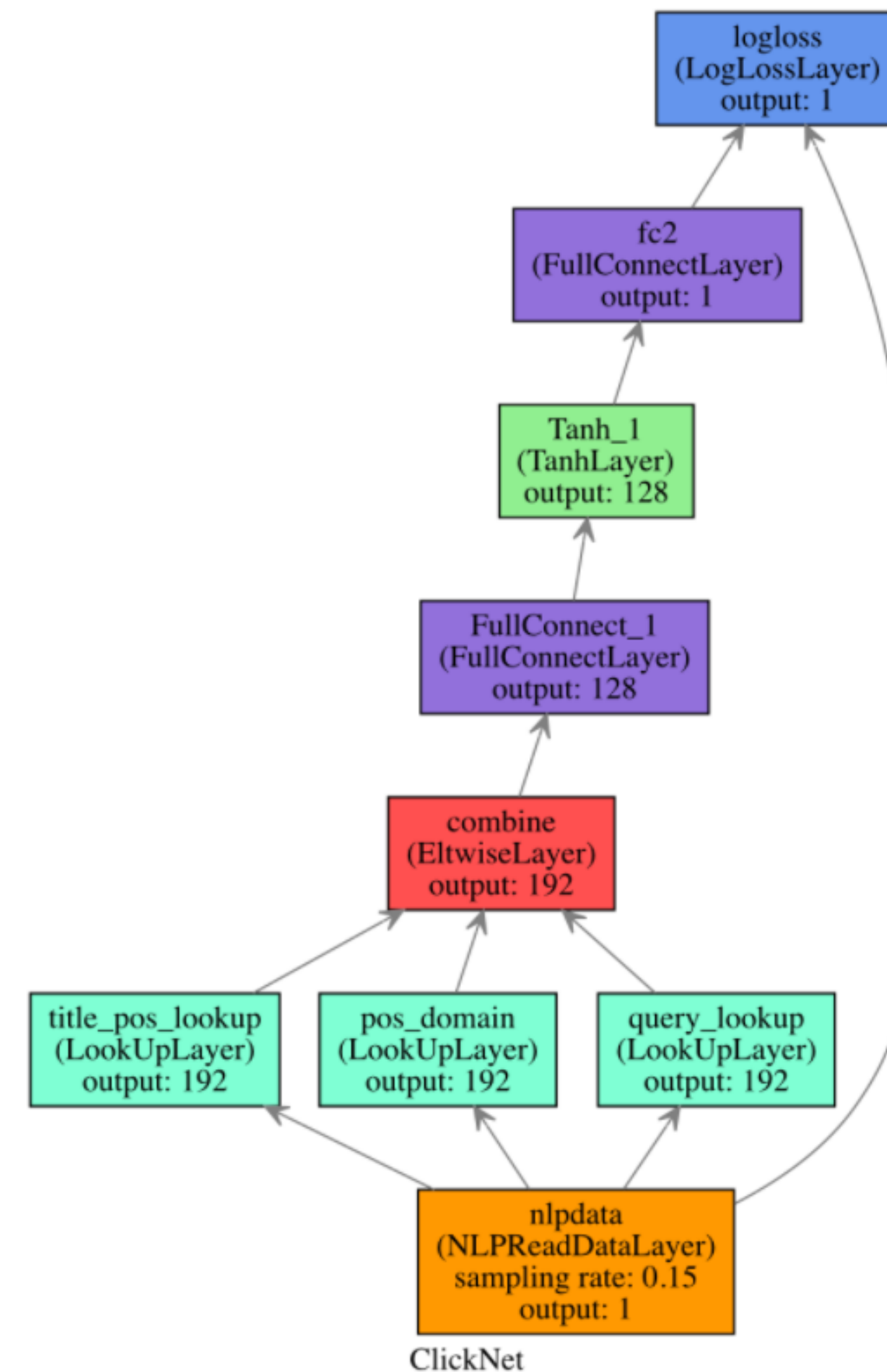
思路

实现

On Ginger（自研的面向文本的深度学习框架）

应用：

ClickNet应用在搜索，广告，酒店，  
旅游等排序系统中



ClickNet on Ginger 语义匹配模型

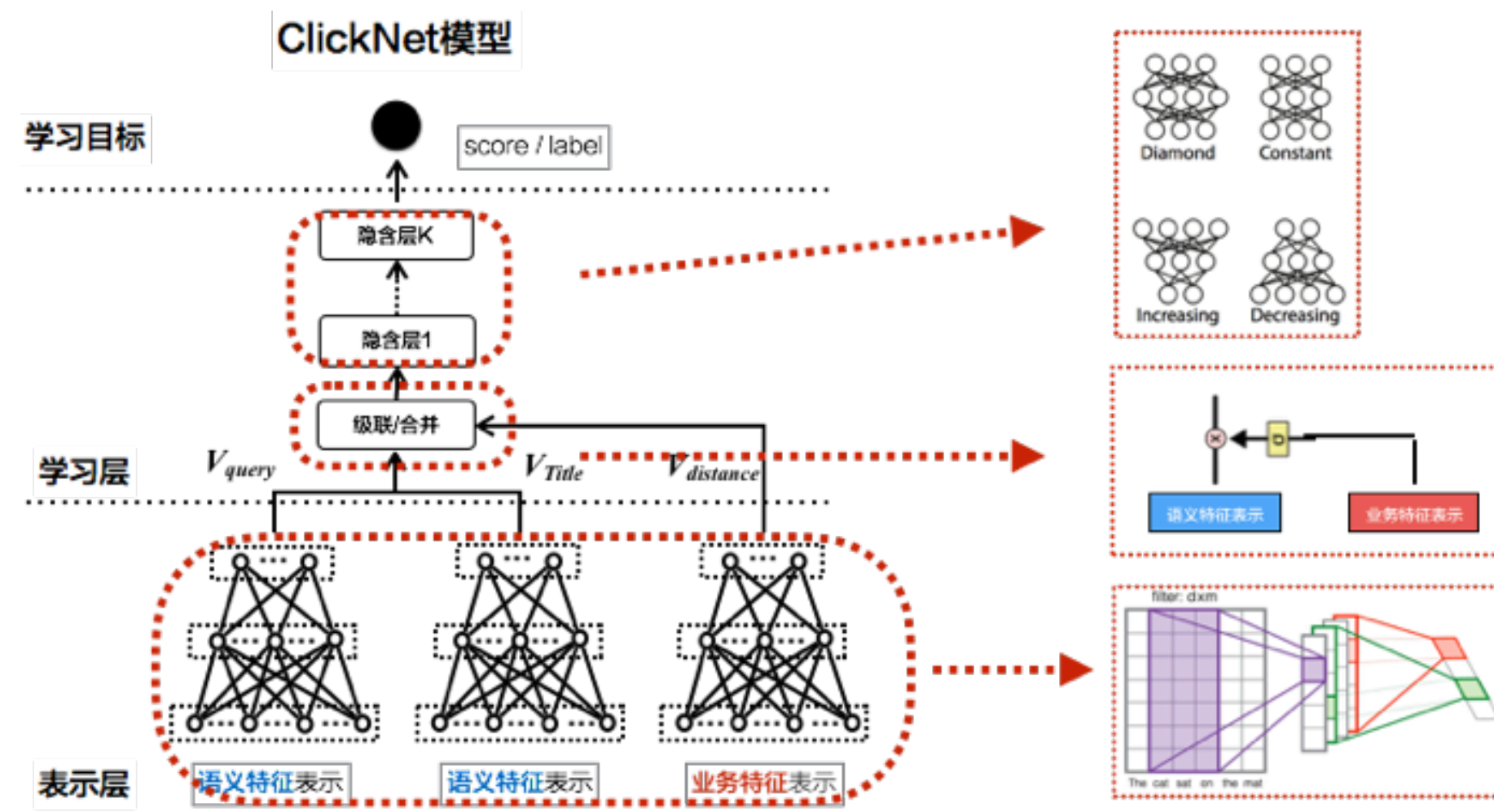


## 案例3 语义匹配模型

### 扩展-ClickNet排序模型

表示层:

加强业务特征即可做CTR模型  
只考虑业务特征可做分类模型



ClickNet 2015

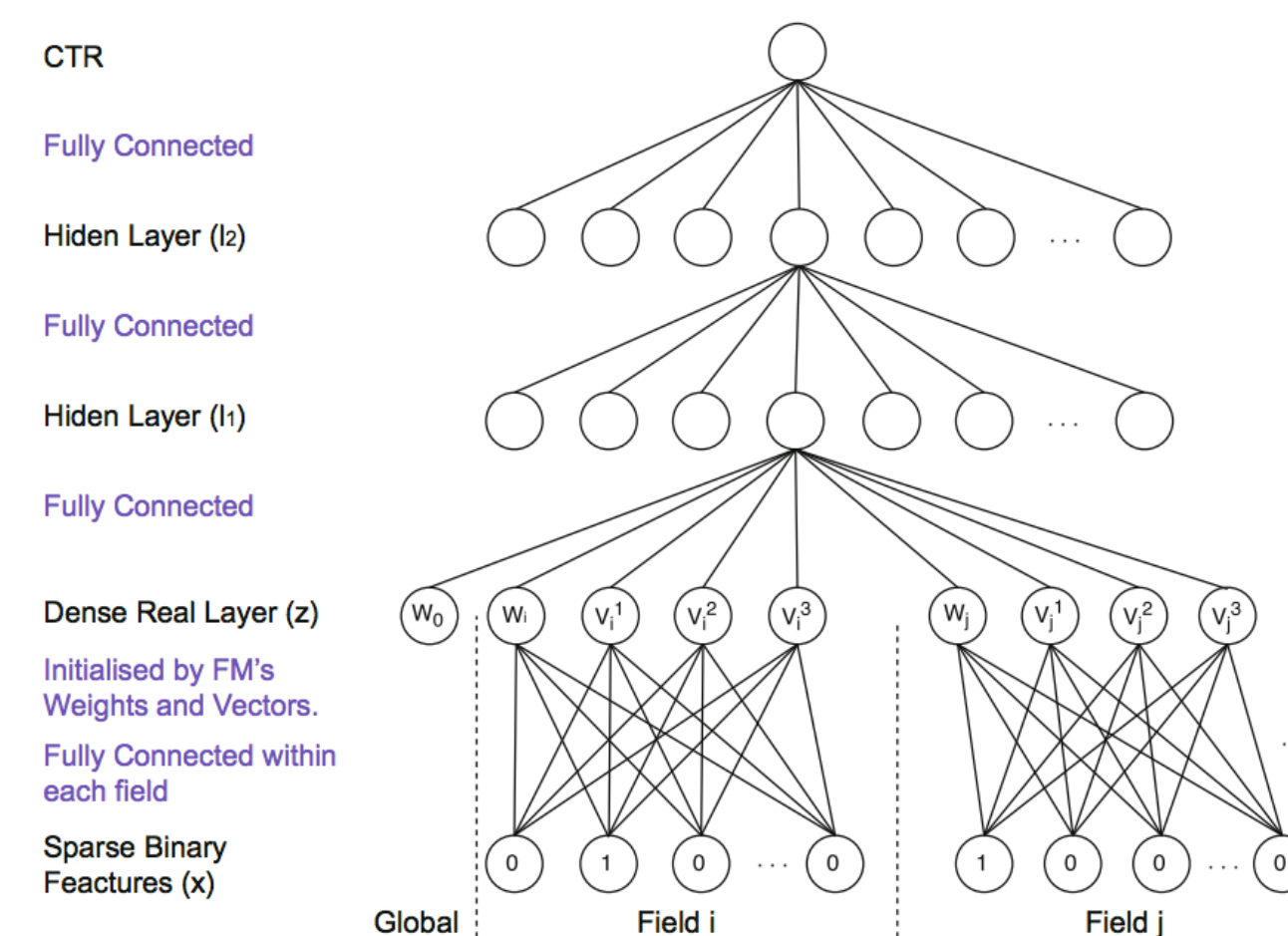
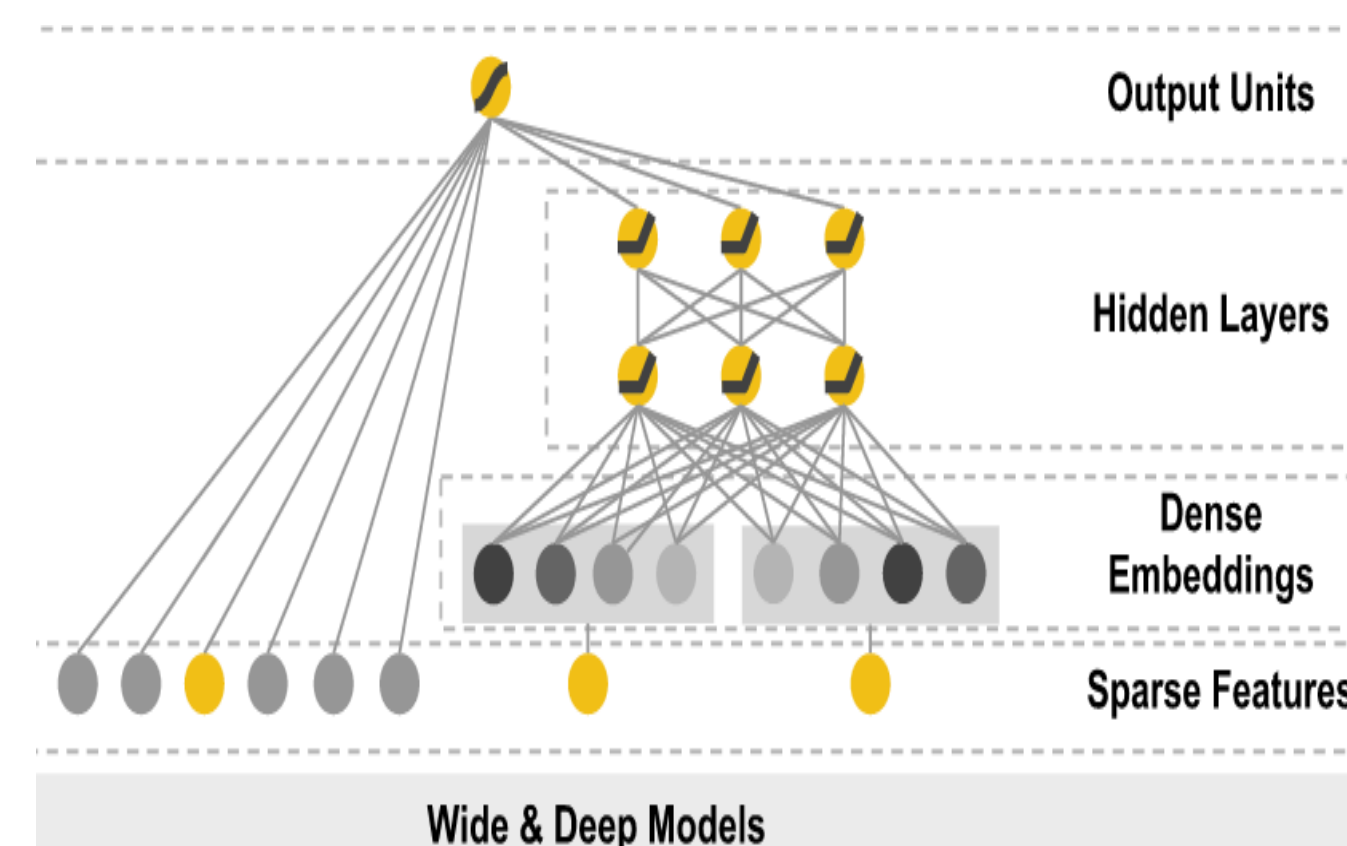


Fig. 1. A 4-layer FNN model structure.

FNN 2016



Google Wide&Deep 2016

# 案例3 语义匹配模型

## 扩展-ClickNet排序模型

Higgs数据和XGBoost和TF实现的ClickNet对比

ClickNet和线性模型对比

模型	数据集	训练速度	测试集AUC
XGBoost	HIGGS	5.7 s/epoch (32 worker*2线程, 2 server)	0.8472
Ginger	Train 1050w样本 Test 50w样本	120 s/epoch (单机24线程)	0.8817
Tensorflow	28维稠密	300 s/epoch (单机GPU+多线程)	0.8787

数据名称	数据大小	特征维度	#nonzeros per instance	离散值/连续值	ginger模型结构 (lookup * hiddens)	训练时间	准确率
train.3 test.3	train: 1243 lines test: 41 lines	21	21 DENSE	连续值	16 * 8	libsvm: 0.1s ginger: 1s/100iters	libsvm: 87.80% ginger: 95.12%
rcv1_train.binary rcv1_test.binary	train: 20242 lines test: 677399 lines	47236	74 SPARSE	连续值	128 * 64 * 32	libsvm: 182.7s liblinear: 0.56s ginger: 84s/100iters	libsvm: 95.14% liblinear: 95.37% ginger: 95.95%
a9a (train) a9a.t (test)	train: 32561 lines test: 16281 lines	123	14 SPARSE	离散值	192 * 128 * 64	libsvm: 53.9s ginger: 103s/100iters	libsvm: 85.07% ginger: 85.31%
kdda (train) kdda.t (test)	train: 8407752 lines test: 540000 lines	20216830	36 SPARSE	连续值	128 * 64	liblinear: 23m38s ginger: 26m/100iters	liblinear: 88.53% ginger: 89.12%

准确率

libsvm: 87.80%

ginger: 95.12%

libsvm: 95.14%

ginger: 95.95%

libsvm: 85.07%

ginger: 85.31%

liblinear: 88.53%

ginger: 89.12%



## 案例4 基于语义表示的Crash在线聚类

场景：

大量Crash需要归类便于处理

难点：

1 如何计算两段Crash相似度

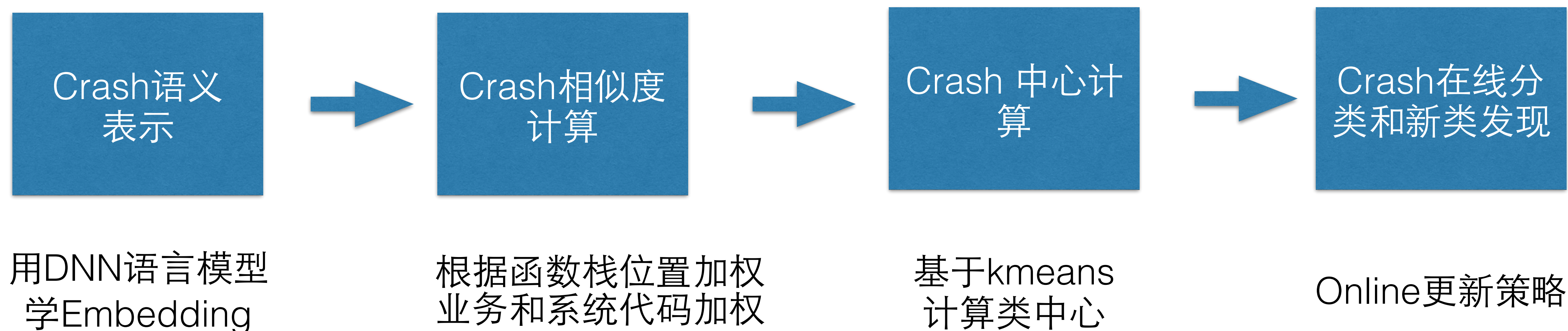
2 类别是动态的，随版本，问题解决情况会变

Thread 0 Crashed:

```
ImageIO      _CGImageReadSessionFinalize(void const*) + 44
CoreFoundation  _CFRelease + 216
ImageIO      IIIOImagePlus::~~IIIOImagePlus() + 168
ImageIO      IIIOImagePlus::~~IIIOImagePlus() + 20
ImageIO      _CGImagePlusFinalize(void const*) + 192
CoreFoundation  _CFRelease + 216
CoreFoundation  __CFBasicHashDrain + 288
CoreFoundation  _CFRelease + 216
ImageIO      IIIOReadPlugin::ReleaseInfo(void*) + 32
ImageIO      IIIOImageProviderInfo::~~IIIOImageProviderInfo() + 156
ImageIO      IIIOImageProviderInfo::~~IIIOImageProviderInfo() + 20
ImageIO      IIIOImageProviderInfo::ReleaseInfo(void*) + 164
CoreGraphics    image_provider_finalize + 148
CoreFoundation  _CFRelease + 216
CoreGraphics    image_finalize + 88
CoreFoundation  _CFRelease + 216
```

## 案例4 基于语义表示的Crash在线聚类

思路

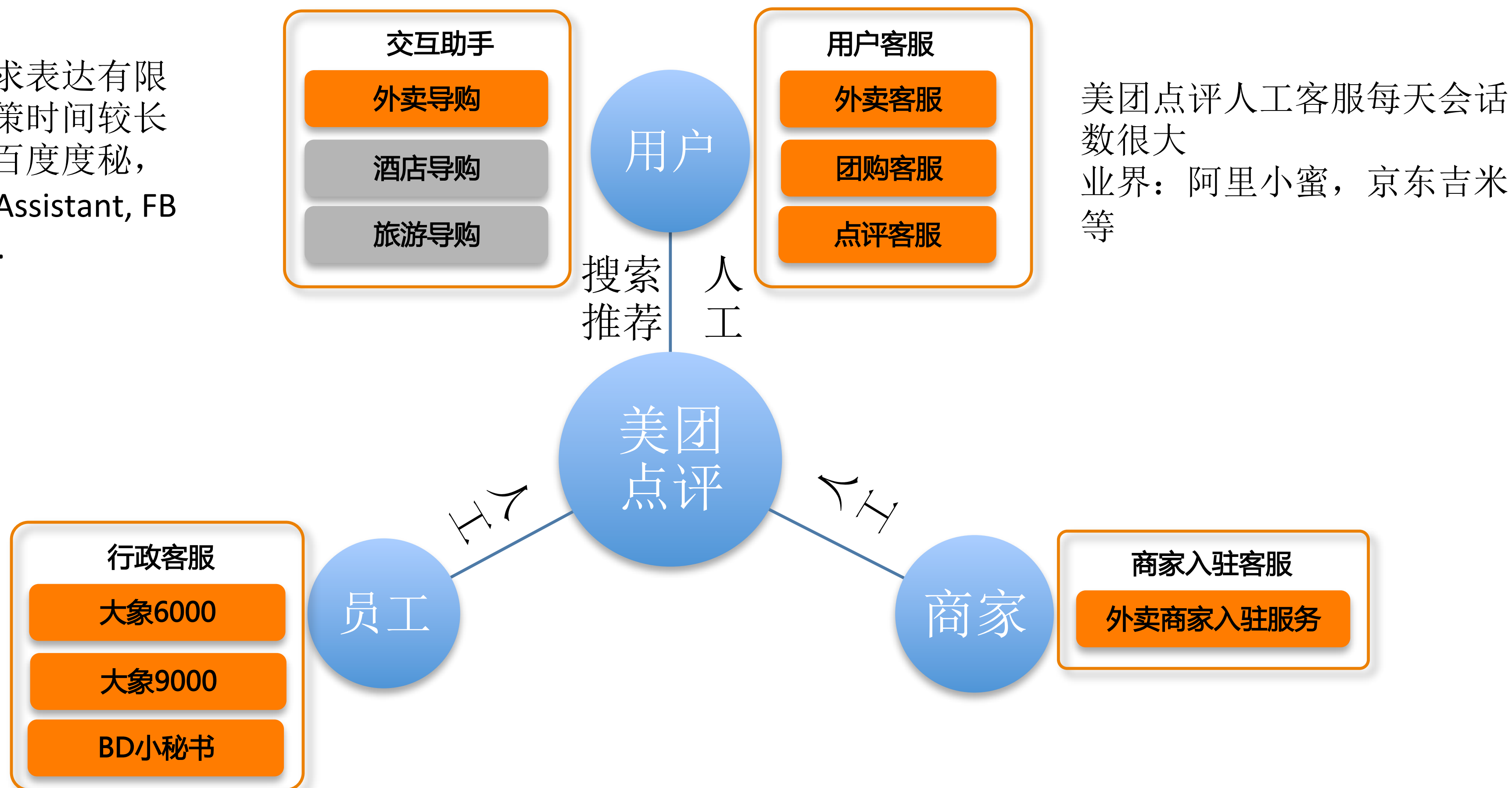


已上线，自动分类和发现新类别

## 案例5 智能客服-小美机器人

### 场景

用户需求表达有限  
消费决策时间较长  
业界：百度度秘，  
Google Assistant, FB  
的“M”..





## 案例5 智能客服-小美机器人

### 小美机器人

人工客服：人力成本高，响应不一定及时

机器人客服：人力成本低，响应快，便于统一监控

类似问题多：

“退款未到账的原因？”

“为什么退款了还没到账”

“退款未到账”



美团app-我的-客服中心-在线客服（小美客服）



## 案例5 智能客服-小美机器人

### 思路

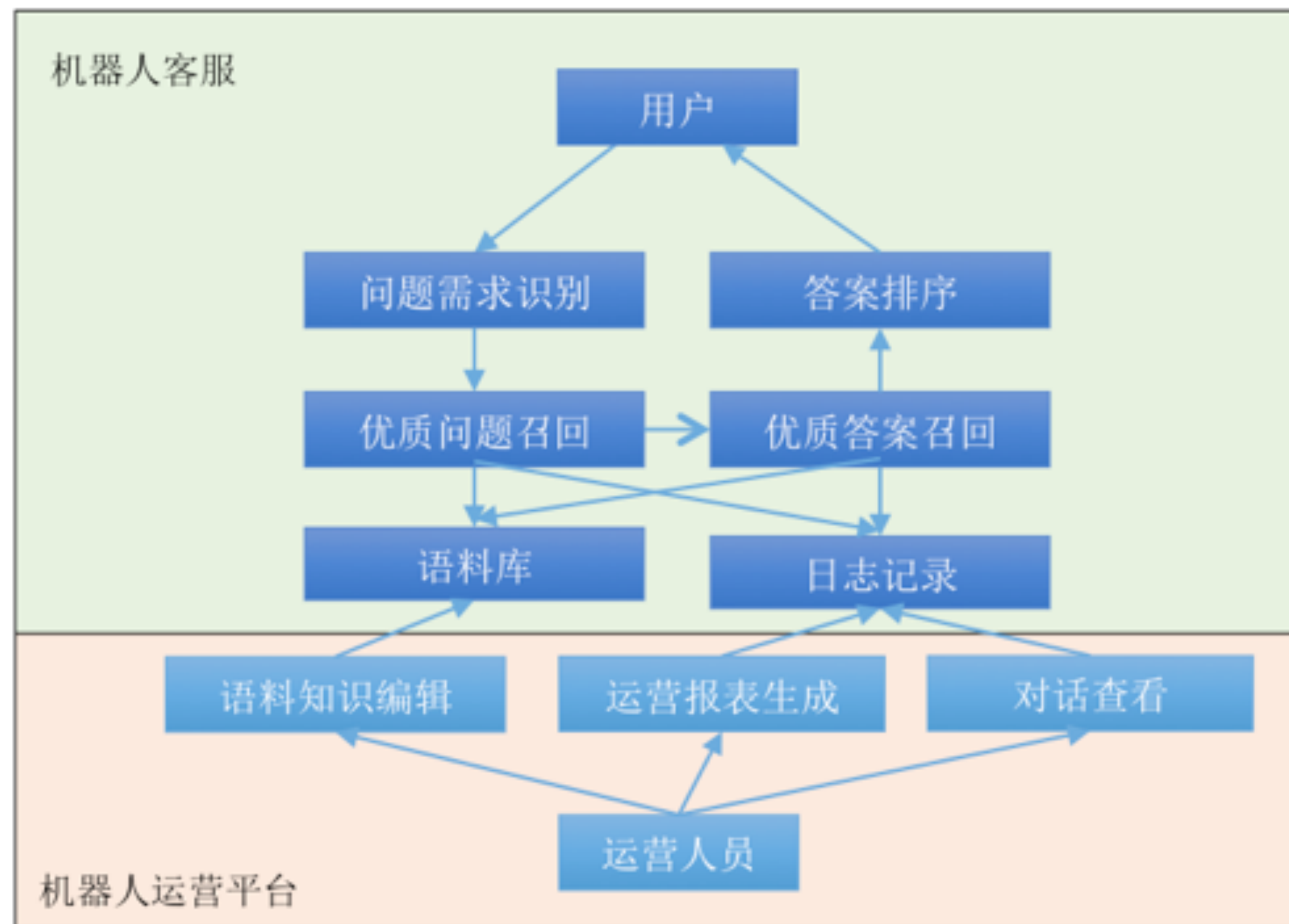
#### 机器人客服

问题需求识别-NLU  
问题相似度计算-语义匹配  
优质问题答案挖掘-DM  
答案排序-Rank

#### 机器人运营平台

知识编辑  
报表生成  
对话查看

解决率超过80%,节省大量人力成本



# 总结

## 01 NLP在美团点评的应用场景

- Query
- UGC评论
- 客服问答
- ...
- 商家标题, 描述, Deal
- 头条内容
- Crash日志

## 02 NLP在美团点评的应用技术

- NLP基础技术
- NLP应用技术

## 03 NLP在美团点评的应用案例

- Query意图识别
- 商家评价摘要
- 语义匹配模型
- 语义表示的Crash聚类
- 客服机器人

# 总结

欢迎加入美团点评智能技术中心！

自然语言处理  
计算机视觉  
用户画像  
深度学习算法  
深度学习框架  
机器学习

.....

邮箱: [nlpml\\_job@meituan.com](mailto:nlpml_job@meituan.com)



Q&A

# Thanks

*Eat better, Live better.*

